F-98RL0111-US

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００２年１１月２８日

出願番号
Application Number: 特願２００２－３４５９８８

[ST. 10/C]: [ＪＰ２００２－３４５９８８]

出願人
Applicant(s): 沖電気工業株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２００３年１１月２８日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫

【書類名】　特許願
【整理番号】　KT000468
【提出日】　平成14年11月28日
【あて先】　特許庁長官　太田　信一郎　殿
【国際特許分類】　G06F 17/28
【発明者】
【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
【氏名】　介弘　達哉
【特許出願人】
【識別番号】　000000295
【氏名又は名称】　沖電気工業株式会社
【代理人】
【識別番号】　100095957
【弁理士】
【氏名又は名称】　亀谷　美明
【電話番号】　03-5919-3808
【選任した代理人】
【識別番号】　100096389
【弁理士】
【氏名又は名称】　金本　哲男
【電話番号】　03-3226-6631
【選任した代理人】
【識別番号】　100101557
【弁理士】
【氏名又は名称】　萩原　康司
【電話番号】　03-3226-6631

【手数料の表示】

【予納台帳番号】　040224

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9707549

【包括委任状番号】　9707550

【包括委任状番号】　9707551

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　複数言語文書の対応付けシステム，複数言語文書の対応付け方法，及びプログラム並びにプログラムを記録した記録媒体

【特許請求の範囲】

【請求項１】　ｎ種（ｎは２以上の自然数）の言語の文書を対応付けるシステムであって，

各言語の文書を単語毎に分割する形態素解析手段と，

前記ｎ種の言語の文書のうちの２種を選択する手段と，

前記選択された２種の言語の文書の評価関数を計算する手段と，

前記２種の言語の文書の評価結果によって，前記ｎ種の言語の文書を対応付ける手段と，

を含むことを特徴とする，複数言語文書の対応付けシステム。

【請求項２】　前記形態素解析手段が，

各言語の文書を文毎に分割する手段と，分割された各文をさらに単語毎に分割する手段とからなることを特徴とする，請求項１に記載の複数言語文書の対応付けシステム。

【請求項３】　前記ｎ種の言語の文書のうちの２種を選択する手段が，

前記ｎ種の言語の文書を任意の順序で並べたときに，ｋ番目とｋ＋１番目（ｋは，１からｎ－１までの自然数）の，ｎ－１通りの組合せを選択することを特徴とする，請求項１または２のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項４】　前記ｎ種の言語の文書のうちの２種を選択する手段が，

ｎ（ｎ－１）／２通りの全ての組合せを選択することを特徴とする，請求項１または２のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項５】　さらに，前記評価関数で計算した結果を保持する計算結果管理手段を含むことを特徴とする，請求項１，２，３，または４のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項６】　前記評価関数が，次式；

$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$

で表されることを特徴とする，請求項１，２，３，４，または５のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

但し，ｈ（ｘ，ｙ）は，評価関数，

ｘは，一方の言語の文，

ｙは，他方の言語の文，

ｆｍ（ｘ，ｙ）は，文ｘと文ｙの中で対応の付いた自立語の数，

ｆｊ（ｘ）は，文ｘ中の自立語の数，

ｆｊ（ｙ）は，文ｙ中の自立語の数，

である。

【請求項７】　さらに，ｎ種のうちのいずれかの３種以上の言語の文書の対応付けに不整合が生じたときに，前記不整合の箇所を表示する手段を含むことを特徴とする，請求項１，２，３，４，５，または６のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項８】　前記評価関数を計算する手段は，前記評価関数の和が最大になるように最適化しながら対応付けを行うことを特徴とする，請求項１，２，３，４，５，６，または７のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項９】　さらに，言語間の類似度データを調べながら，対応付けの正解率の高い言語対を指示する手段を含むことを特徴とする，請求項１，２，３，４，５，６，７，または８のうちのいずれか１項に記載の複数言語文書の対応付けシステム。

【請求項１０】　ｎ種（ｎは２以上の自然数）の言語の文書を対応付ける方法であって，

各言語の文書を単語毎に分割する形態素解析段階と，

前記ｎ種の言語の文書のうちの２種を選択する段階と，

前記選択された２種の言語文書の評価関数を計算する段階と，

前記２種の言語の文書の評価結果によって，ｎ種の言語の文書を対応付ける段階と，

を含むことを特徴とする，複数言語文書の対応付け方法。

【請求項１１】　前記形態素解析段階が，

各言語の文書を文毎に分割する段階と，分割された各文をさらに単語毎に分割する段階とからなることを特徴とする，請求項１０に記載の複数言語文書の対応付け方法。

【請求項１２】　前記ｎ種の言語の文書のうちの２種を選択する段階が，

前記ｎ種の言語の文書を任意の順序で並べたときに，ｋ番目とｋ＋１番目（ｋは，１からｎ－１までの自然数）の，ｎ－１通りの組合せを選択することを特徴とする，請求項１０または１１のうちのいずれか１項に記載の複数言語文書の対応付け方法。

【請求項１３】　前記ｎ種の言語の文書のうちの２種を選択する段階が，

ｎ（ｎ－１）／２通りの全ての組合せを選択することを特徴とする，請求項１０または１１のうちのいずれか１項に記載の複数言語文書の対応付け方法。

【請求項１４】　さらに，前記評価関数で計算した結果を保持する計算結果管理段階を含むことを特徴とする，請求項１０，１１，１２，または１３のうちのいずれか１項に記載の複数言語文書の対応付け方法。

【請求項１５】　前記評価関数が，次式；

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

で表されることを特徴とする，請求項１０，１１，１２，１３，または１４のうちのいずれか１項に記載の複数言語文書の対応付け方法。

但し，ｈ（ｘ，ｙ）は，評価関数，

ｘは，一方の言語の文，

ｙは，他方の言語の文，

f_m（ｘ，ｙ）は，文ｘと文ｙの中で対応のついた自立語の数，

f_j（ｘ）は，文ｘ中の自立語の数，

f_j（ｙ）は，文ｙ中の自立語の数，

である。

【請求項１６】　さらに，ｎ種のうちのいずれかの３種以上の言語の文書の対応付けに不整合が生じたときに，前記不整合の箇所を表示する段階を含むことを特徴とする，請求項１０，１１，１２，１３，１４，または１５のうちのいず

れか１項に記載の複数言語文書の対応付け方法。

【請求項１７】　前記評価関数を計算する段階は，前記評価関数の和が最大になるように最適化しながら対応付けを行うことを特徴とする，請求項１０，１１，１２，１３，１４，１５，または１６のうちのいずれか１項に記載の複数言語文書の対応付け方法。

【請求項１８】　さらに，言語間の類似度データを調べながら，対応付けの正解率の高い言語対を指示する段階を含むことを特徴とする，請求項１０，１１，１２，１３，１４，１５，１６，または１７のうちのいずれか１項に記載の複数言語文書の対応付け方法。

【請求項１９】　コンピュータに，請求項１０，１１，１２，１３，１４，１５，１６，１７，または１８のうちのいずれか１項に記載の複数言語文書の対応付け方法を行わせるステップを記述したことを特徴とする，プログラム。

【請求項２０】　コンピュータに，請求項１０，１１，１２，１３，１４，１５，１６，１７，または１８のうちのいずれか１項に記載の複数言語文書の対応付け方法を行わせるプログラムが記録されていることを特徴とする，記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は，複数の言語で構成される文書間の文書対応付けシステムにかかり，特に，２言語以上で記述された対訳文書の，文の対応付けを行う複数言語文書の対応付けシステム，複数言語文書の対応付け方法，この方法を行わせるプログラム，及びこのプログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】

海外に輸出される製品のマニュアルなどのように，複数の言語で同じ内容の文書を記述する場合が増えている。このような複数の言語文書の対訳の正確性を評価，担保等するため，これらの文の対応付けを行う需要も増えている。非特許文献１は，対訳文書の文の対応付けを，対訳辞書を利用したダイナミックプログラ

ミングで行う方法が記載されている。

【０００３】

非特許文献１によれば，対応付けを行うには，文書を１文毎に区切り，さらにその文の形態素解析を行って，単語毎に分割する。そして，これらの単語の中から自立語を取り出し，対訳辞書を用いてそれぞれの文の中の自立語がどの程度対応しているか（どの程度意味内容が一致しているか）によって対応付けを評価する。評価では，例えば以下のような式を用いる。

【０００４】

【数１】

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

【０００５】

ここで

ｈ（ｘ，ｙ）は，評価関数，

ｘは，原文中の文（複数文の場合もある），

ｙは，訳文中の文（複数文の場合もある），

f_m（ｘ，ｙ）は，文ｘと文ｙの中で対応の付いた自立語の数，

f_j（ｘ）は，文ｘ中の自立語の数，

f_j（ｙ）は，文ｙ中の自立語の数，

である。

【０００６】

このような式による評価を行えば，文書の対応の割合が大きいほど評価関数ｈ（ｘ，ｙ）の値は大きくなり（最大；１），逆は小さくなる（最小；０）。この評価関数を文の先頭から調べていき，評価関数の和が最も大きくなる組合せを，対応付け問題の解とする。

【０００７】

【非特許文献１】

宇津呂　武仁，松本　裕治　共著「対訳辞書及び統計情報を用いた二言語対訳テキスト照合」（「コンピュータソフトウェア」岩波書店　vol.12 No.5 Sep.1995 p.12(414)-p.21(423)）

【０００８】

【発明が解決しようとする課題】

しかしながら，上記方法では，通常の２言語の対訳文書の文の対応付けを，３言語以上の文書の文の対応付けに適用する場合に，

・複数の辞書を使用するため，システムにかなりの量の記録領域を必要とする。

・評価の処理に時間がかかる。

・全ての言語間で，各言語対の対応の整合性をとるのが困難である。

などの問題がある。

【０００９】

また，２言語の対訳文書の対応付けに関しても，高精度での対応を自動的に付けるのは難しく，対応付けの結果を見ながらの人の手によるチェックや修正が必要であり，その作業時間が問題となっている。

【００１０】

本発明は，従来の複数言語文書の対応付けシステムが有する上記問題点に鑑みてなされたものである。そして，本発明の目的は，英語－日本語－ドイツ語など，複数の言語でそれぞれ構成される文書間の文の対応付けを効率良く行うための，新規かつ改良された複数言語文書の対応付けシステム，及び複数言語文書の対応付け方法を提供することにある。

【００１１】

【課題を解決するための手段】

上記課題を解決するための本発明の複数言語文書の対応付けシステムは，ｎ種（ｎは２以上）の言語の文書を対応付けるシステムである。そして，各言語の文書を単語毎に分割する形態素解析手段と，ｎ種の言語の文書のうちの２種を選択する手段と，選択された２種の言語文書の評価関数を計算する手段と，評価結果に応じてｎ種の言語の文書を対応付ける手段とから構成される。

【００１２】

ここで，各言語の文書を単語毎に分割する形態素解析手段は，各言語の文書を文毎に分割する手段と，分割された各文をさらに単語毎に分割する手段とからなってもよい。

【００１３】

【発明の実施の形態】

以下に添付図面を参照しながら，本発明にかかる複数言語文書の対応付けシステム，複数言語文書の対応付け方法の好適な実施の形態について詳細に説明する。

【００１４】

（第１の実施の形態）

図１は，第１の実施の形態にかかる複数言語文書の対応付けシステム１００の構成を示す説明図である。複数言語文書の対応付けシステム１００は，図１に示したように，文分割手段１０５と，形態素解析手段１０６と，評価関数計算手段１０７と，計算結果管理手段１０８と，対訳辞書データベース１０９により構成されている。この例では，各言語のファイル１０１～１０４が入力されて，対応タグ付きファイル１１０～１１３が出力される。以下，各構成要素につき詳細に説明する。

【００１５】

英語ファイル１０１は，英語で記述された文書ファイル，日本語ファイル１０２は，日本語で記述された文書ファイル，ドイツ語ファイル１０３は，ドイツ語で記述された文書ファイル，中国語ファイル１０４は，中国語で記述された文書ファイルである。上記４つのファイルはそれぞれ同じ内容のことを述べており，それぞれが対訳形式になっている。

【００１６】

文分割手段１０５は，文書ファイルを１文毎に分割する。例えば，英文であればピリオド「．」，日本文なら句点「。」などで分割する。形態素解析手段１０６は，形態素解析処理を行い，文を単語毎に分割する。文分割手段１０５及び形態素解析手段１０６は，既存のものを適用できる。

【００１７】

評価関数計算手段１０７は，最適な対応付けを見つけるために，与えられた評価関数を計算する。例えば，評価関数は，次式；

$$h(x, y) = 2 \times f_m(x, y) / (f_j(x) + f_j(y))$$

で表される。ここで，ｈ（ｘ，ｙ）は，評価関数であり，ｘは，一方の言語の文（原文）であり，ｙは，他方の言語の文（訳文）であり，f_m（ｘ，ｙ）は，文ｘと文ｙの中で対応の付いた自立語の数であり，f_j（ｘ）は，文ｘ中の自立語の数であり，f_j（ｙ）は，文ｙ中の自立語の数である。

【００１８】

計算結果管理手段１０８は，評価関数計算手段が計算した結果を保持し，既出の評価関数計算が再び到来したときに保持した結果を出力し，同じ計算を何度も行わないようにする。

【００１９】

対訳辞書データベース１０９は，対応付けをするための原文の単語を引くと訳文の語が１つまたは複数あるような辞書である。例えば，原文が英語，訳文が日本語の場合，英和辞典に相当する。

【００２０】

対応タグ付英語ファイル１１０は，英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル１１１，対応タグ付きドイツ語ファイル１１２，及び対応タグ付中国語ファイル１１３も同様に，元のファイルに文の対応を示すためのタグを付与したものである。

【００２１】

本実施の形態にかかる複数言語文書の対応付けシステム１００は，以上のように構成されている。次に，図２を参照しながら，複数言語文書の対応付けシステム１００の動作を説明する。

【００２２】

図２は，第１の実施の形態の複数言語文書の対応付けシステム１００の動作を示すフローチャートである。ステップＳ１０では文分割手段によって一方の文（原文）ファイルと他方の文（訳文）ファイルの文分割を行う。そして，対応付けをどこまで行ったかを示すカウンタＮを，０にセットする。

ステップＳ１１では，カウンタＮをインクリメント（＋１）する。

ステップＳ１２では，対応付けを行う言語の数がカウンタＮと等しいかどうかを比較する。もし等しければ，ステップＳ１７に行く。

【００２３】

ステップＳ１３では，対応付けを行う言語をＮ番目とＮ＋１番目にセットする。

ステップＳ１４では，評価関数計算手段がセットされた言語に対して文の対応付けを行う。

ステップＳ１５では，対応付けを行った結果に対して，対応する文同士に双方向リンクを張る。

【００２４】

ステップＳ１６では，２対１，３対１などの複数文の対応になった文に対してマーク付けを行う。これらのマーク付けされた文の組は，次に対応付けを行う場合はそれを主文とみなして処理する。

ステップＳ１７では，対応付けを行っていない言語同士の文に対して，他の言語同士の対応付け結果を利用して，リンクを張る。

【００２５】

以上の処理を，図１の４言語の対応付けを行う場合に関して説明する。この例では，英語が１番目，日本語が２番目，ドイツ語が３番目，中国語が４番目の言語に相当する。

【００２６】

まず，４つの言語（ｎ＝４）それぞれを文分割手段によって一文毎に分割する。次に，文の対応付けを行う。英語と日本語の対応付けを英日対訳辞書を使って，日本語とドイツ語の対応付けを日独対訳辞書を使って，ドイツ語と中国語の対応付けを独中対訳辞書を使ってそれぞれ行う。これにより，日本語－英語間，日本語－ドイツ語間，ドイツ語－中国語間の（ｎ－１）通りの文同士のリンクが生成される。

【００２７】

さらに，対応のついていない言語同士（ここでは，日本語－中国語，英語－ドイツ語，英語－中国語）の文のリンクを張ることによって，すべての言語間の文の対応をとることができる。

【００２８】

以上説明したように，本実施の形態によれば，対応付けの精度は多少落ちるが，少ない記録容量で時間もあまりかからずに効率良く文の対応をとることができる。

【００２９】

（第２の実施の形態）

図３に，第２の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル２０１は，英語で記述された文書ファイル，日本語ファイル２０２は，日本語で記述された文書ファイル，ドイツ語ファイル２０３は，ドイツ語で記述された文書ファイル，中国語ファイル２０４は，中国語で記述された文書ファイルである。上記４つのファイルは，それぞれ同じ内容が記述されており，それぞれが対訳形式になっている。

【００３０】

文分割手段２０５は，文書ファイルを１文毎に分割する。英文であればピリオド「．」，日本文なら句点「。」などで分割する。形態素解析手段２０６は，形態素解析処理を行い，文を単語毎に分割する。文分割手段２０５及び形態素解析手段２０６は，既存のものを適用できる。評価関数計算手段２０７は，最適な対応付けを見つけるために，与えられた評価関数を計算する。評価関数は，例えば第１の実施の形態で示したものが適用できる。

【００３１】

計算結果管理手段２０８は，評価関数計算手段が計算した結果を保持し，既出の評価関数計算が再び到来したときに保持した結果を出力し，同じ計算を何度も行わないようにする。対訳辞書データベース２０９は，対応付けをするための辞書で，原文の単語を引くと訳文の語が１つまたは複数あるような辞書である。原文が英語，訳文が日本語の場合，英和辞典に相当する。

【００３２】

対応タグ付英語ファイル２１０は，英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル２１１，対応タグ付きドイツ語ファイル２１２，及び対応タグ付中国語ファイル２１３も同様に，元のファイルに文の対応を示すためのタグを付与したものである。

【００３３】

相違箇所表示手段２２０は，対応付け結果に不整合があった場合に，その不整合箇所を表示し，ユーザに修正させる機能をもつ。不整合とは，例えば，英語の文Ｅｎと日本語のＪｎが対応していて，日本語の文Ｊｎとドイツ語の文Ｄｎが対応しているときに，英語とドイツ語の対応結果をみると，英文Ｅｎとドイツ文Ｄｎとが対応していないような場合である。

【００３４】

図４は，本実施の形態の複数言語文書の対応付けシステム２００の動作を示すフローチャートである。

ステップＳ２０では，文分割手段によって一方の文（原文）ファイルと他方の文（訳文）ファイルの文分割を行う。そして，対応付けをどこまで行ったかを示すカウンタＮとＭを，１にセットする。

ステップＳ２１では，対応付けを行う言語の数がカウンタＮと等しいかどうかを比較する。もし等しければ，ステップＳ２７に行く。

ステップＳ２２では，カウンタＭをインクリメントし，Ｎの値をＭ＋１にする。

ステップＳ２３では，対応付けを行う言語の数がカウンタＭと等しいかどうかを比較する。もし等しければ，Ｓ２８に行く。

【００３５】

ステップＳ２４では，対応付けを行う言語をＭ番目とＮ番目にセットする。

ステップＳ２５では，評価関数計算手段がセットされた言語に対して文の対応付けを行う。

ステップＳ２６では，対応付けを行った結果に対して，対応する文同士に双方向リンクを張る。

ステップＳ２７では，Ｎをインクリメントする。

ステップＳ２８では，文の対応に不整合がある部分を表示しユーザに修正させる。

ステップＳ２９では，ユーザの修正に応じて，対応付けのリンクを張り直す。

このようにして，ｎ種の言語の文に対して，全ての組合せ（この例では，言語

の種類ｎ＝４で，ｎ（ｎ－１）／２＝６通り）の対応付けを行う。

【００３６】

以上説明したように，本実施の形態によれば，ユーザが修正することが必須であるが，高精度の対応付けが効率良く実現できる。

【００３７】

（第３の実施の形態）

図５に，第３の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル３０１は，英語で記述された文書ファイル，日本語ファイル３０２は，日本語で記述された文書ファイル，ドイツ語ファイル３０３は，ドイツ語で記述された文書ファイル，中国語ファイル３０４は，中国語で記述された文書ファイルである。上記４つのファイルは，それぞれ同じ内容が記述されており，それぞれが対訳形式になっている。

【００３８】

文分割手段３０５は，文書ファイルを１文毎に分割する。英文であればピリオド「．」，日本文なら句点「。」などで分割する。形態素解析手段３０６は，形態素解析処理を行い，文を単語毎に分割する。文分割手段３０５及び形態素解析手段３０６は，既存のものを適用できる。評価関数計算手段３０７は，最適な対応付けを見つけるために，与えられた評価関数を計算する。評価関数は，例えば第１の実施の形態で示したものが適用できる。

【００３９】

計算結果管理手段３０８は，評価関数計算手段が計算した結果を保持し，既出の評価関数計算が再び到来したときに保持した結果を出力し，同じ計算を何度も行わないようにする。

【００４０】

対訳辞書データベース３０９は，対応付けをするための辞書で，原文の単語を引くと訳文の語が１つまたは複数あるような辞書である。原文が英語，訳文が日本語の場合英和辞典に相当する。

【００４１】

対応タグ付英語ファイル３１０は，英語ファイルに他の文書のどの文に対応し

ているかを示すタグを付与したものである。対応タグ付き日本語ファイル３１１，対応タグ付きドイツ語ファイル３１２，及び対応タグ付中国語ファイル３１３も同様に，元のファイルに文の対応を示すためのタグを付与したものである。

【００４２】

図６は，本実施の形態の複数言語文書の対応付けシステム３００の動作を示すフローチャートである。

ステップＳ３０では，文分割手段によって一方の文（原文）ファイルと他方の文（訳文）ファイルの文分割を行う。そして，対応付けをどこまで行ったかを示すカウンタＮとＭを，１にセットする。

ステップＳ３１では，対応付けを行う言語の数がカウンタＮと等しいかどうかを比較する。もし等しければ，ステップＳ３７に行く。

【００４３】

ステップＳ３２では，カウンタＭをインクリメントし，Ｎの値をＭ＋１にする。

ステップＳ３３では，対応付けを行う言語の数がカウンタＭと等しいかどうかを比較する。もし等しければ，ステップＳ３７に行く。

【００４４】

ステップＳ３４では，対応付けを行う言語をＭ番目とＮ番目にセットする。

ステップＳ３５では，評価関数計算手段がセットされた言語に対して評価関数を計算する。

ステップＳ３６では，Ｎをインクリメントする。

ステップＳ３７では，対応付けのポイントの和が最も大きくなるような文の組を選択する。

ステップＳ３８では，対応する文同士に双方向リンクを張る。

【００４５】

以上の処理を，図５の４言語（ｎ＝４）の対応付けを行う場合に関して説明する。例では，英語が１番目，日本語が２番目，ドイツ語が３番目，中国語が４番目の言語に相当する。

【００４６】

まず，４つの言語それぞれを文分割手段によって一文毎に分割する。次に，すべての文書の組の評価関数を計算する。この場合，英語－日本語，英語－ドイツ語，英語－中国語，日本語－ドイツ語，日本語－中国語，ドイツ語－中国語の６つの評価関数を計算する。

【００４７】

次に，対応付けポイントの和が最も大きくなるように対応をとっていく。この対応は４言語まとめて同時に行われる。例えば，英文１文，日本文１文，ドイツ文２文，中国文１文の評価ポイントは，英文と日本文の１文対１文，英文とドイツ文の１文対２文，英文と中国文の１文対１文，日本文とドイツ文の１文対２文，日本文と中国文の１文対１文，ドイツ文と中国文の２文対１文，の評価ポイントの和となる。この計算を続け，評価ポイントの和の和が最も大きくなったものを対応付けの正解とする。

【００４８】

以上説明したように，本実施の形態によれば，時間はかなりかかるが高精度の対応付けが効率良く実現できる。

【００４９】

（第４の実施の形態）

図７に，第４の実施の形態の複数言語文書の対応付けシステムの構成を示す。英語ファイル４０１は英語で記述された文書ファイル，日本語ファイル４０２は日本語で記述された文書ファイル，ドイツ語ファイル４０３はドイツ語で記述された文書ファイル，中国語ファイル４０４は中国語で記述された文書ファイルである。上記４つのファイルはそれぞれ同じ内容が記述されており，それぞれが対訳形式になっている。

【００５０】

文分割手段４０５は，文書ファイルを１文毎に分割する。英文であればピリオド「．」，日本文なら句点「。」などで分割する。形態素解析手段４０６は，形態素解析処理を行い，文を単語毎に分割する。文分割手段４０５及び形態素解析手段４０６は，既存のものを適用できる。評価関数計算手段４０７は，最適な対応付けを見つけるために，与えられた評価関数を計算する。評価関数は，例えば

第１の実施の形態で示したものが適用できる。

【００５１】

計算結果管理手段４０８は，評価関数計算手段が計算した結果を保持し，既出の評価関数計算が再び到来したときに保持した結果を出力し，同じ計算を何度も行わないようにする。対訳辞書データベース４０９は，対応付けをするための辞書で，原文の単語を引くと訳文の語が１つまたは複数あるような辞書である。原文が英語，訳文が日本語の場合英和辞典に相当する。

【００５２】

対応タグ付英語ファイル４１０は，英語ファイルに他の文書のどの文に対応しているかを示すタグを付与したものである。対応タグ付き日本語ファイル４１１，対応タグ付きドイツ語ファイル４１２，及び対応タグ付中国語ファイル４１３も同様に，元のファイルに文の対応を示すためのタグを付与したものである。

【００５３】

言語類似度データ４２０は，言語同士の文法などがどれだけ似ているかを数値化したものである。類似度が高いほど文の対応付けの程度も向上する。それぞれの言語対の類似度の値が，例えば表形式などで記録されている。

【００５４】

図８は，本実施の形態の複数言語文書の対応付けシステム４００の動作を示すフローチャートである。

ステップＳ４０では，文分割手段によって一方の文ファイルと他方の文ファイルの文分割を行う。

対応付けをどこまで行ったかを示すカウンタＮを，０にセットする。

ステップＳ４１では，カウンタＮをインクリメントする。

ステップＳ４２では，対応付けを行う言語の数がカウンタＮと等しいかどうかを比較する。もし等しければ，終了する。

【００５５】

ステップＳ４３では，言語類似度が最も高く，まだ選択されていない言語対を選択し，選択済みのマークをつけておく。

ステップＳ４４では，言語対に文対応のリンクが張られているかどうかを調べ

る。リンクがすでに張られていれば，ステップＳ４３に行く。

【００５６】

ステップＳ４５では，評価関数計算手段が選択された言語に対して文の対応付けを行う。

ステップＳ４６では，対応付けを行った結果に対して，対応する文同士に双方向リンクを張る。

ステップＳ４７では，２対１，３対１などの複数文の対応になった文に対してマーク付けを行う。これらのマーク付けされた文の組は次に対応付けを行う場合はそれを１文とみなして処理する。

ステップＳ４８では，間接的に対応のついた言語に対してリンクを張る。例えば，英語－日本語，英語－ドイツ語の対応がとれたとすると，日本語－ドイツ語間にも文対応のリンクを張る。

【００５７】

以上説明したように，本実施の形態によれば，言語類似度データを用意する必要があるが，高速に精度の高い対応付けが効率良く実現できる。

【００５８】

上記の４つの実施の形態の速度，精度，使用する記録容量を比較すると，表１のようになる。表１において，「◎」は優良，「○」は良好，「△」は普通である。

【００５９】

【表１】

実施の形態	速度	精度	記録容量	その他
1	◎	△	◎	
2	○	◎	△	ユーザの修正が必須
3	△	◎	△	
4	◎	○	○	言語類似度データが必要

【００６０】

以上，添付図面を参照しながら本発明にかかる複数言語文書の対応付けシステ

ム，及び複数言語文書の対応付け方法の好適な実施形態について説明したが，本発明はかかる例に限定されない。当業者であれば，特許請求の範囲に記載された技術的思想の範疇内において各種の変更例または修正例に想到し得ることは明らかであり，それらについても当然に本発明の技術的範囲に属するものと了解される。

【００６１】

例えば，上記第１～第４実施の形態では，英語，日本語，ドイツ語，中国語の対応付けを示したが，対訳辞書を変えることによって，どんな言語同士の対応もとることができる。また，４言語（ｎ＝４）の例を示したが，２言語以上であれば何言語の対応付けにも対応できる。第２，第３の実施の形態では言語数が増えてくると処理時間が非常に遅くなるおそれがあるが，計算する対応組の数を減らすことによって対応できる。

【００６２】

なお，本発明の複数言語文書の対応付け方法は，プログラムに記述することもでき，本発明の複数言語文書の対応付け方法を記述したプログラムは，記録媒体に記録することができる。

【００６３】

【発明の効果】

以上説明したように，本発明によれば，複数の言語で構成される文書間の文の対応付けを効率良く行う複数言語文書の対応付けシステムが提供できた。

【図面の簡単な説明】

【図１】

第１の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図２】

図１の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図３】

第２の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図４】

図３の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図５】

第３の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図６】

図５の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【図７】

第４の実施の形態にかかる複数言語文書の対応付けシステムの構成を示す説明図である。

【図８】

図７の複数言語文書の対応付けシステムの動作を示すフローチャートである。

【符号の説明】

１００，２００，３００，４００	複数言語文書の対応付けシステム
１０１，２０１，３０１，４０１	英語ファイル
１０２，２０２，３０２，４０２	日本語ファイル
１０３，２０３，３０３，４０３	ドイツ語ファイル
１０４，２０４，３０４，４０４	中国語ファイル
１０５，２０５，３０５，４０５	文分割手段
１０６，２０６，３０６，４０６	形態素解析手段
１０７，２０７，３０７，４０７	評価関数計算手段
１０８，２０８，３０８，４０７	計算結果管理手段
１０９，２０９，３０９，４０９	対訳辞書データベース
１１０，２１０，３１０，４１０	対応タグ付英語ファイル
１１１，２１１，３１１，４１１	対応タグ付日本語ファイル
１１２，２１２，３１２，４１２	対応タグ付独語ファイル
１１３，２１３，３１３，４１３	対応タグ付中国語ファイル
１１４，２１４，３１４，４１４	英日対訳辞書
１１５，２１５，３１５，４１５	日独対訳辞書

１１６，２１６，３１６，４１６　独中対訳辞書

２１７，３１７，４１７　英独対訳辞書

２１８，３１８，４１８　英中対訳辞書

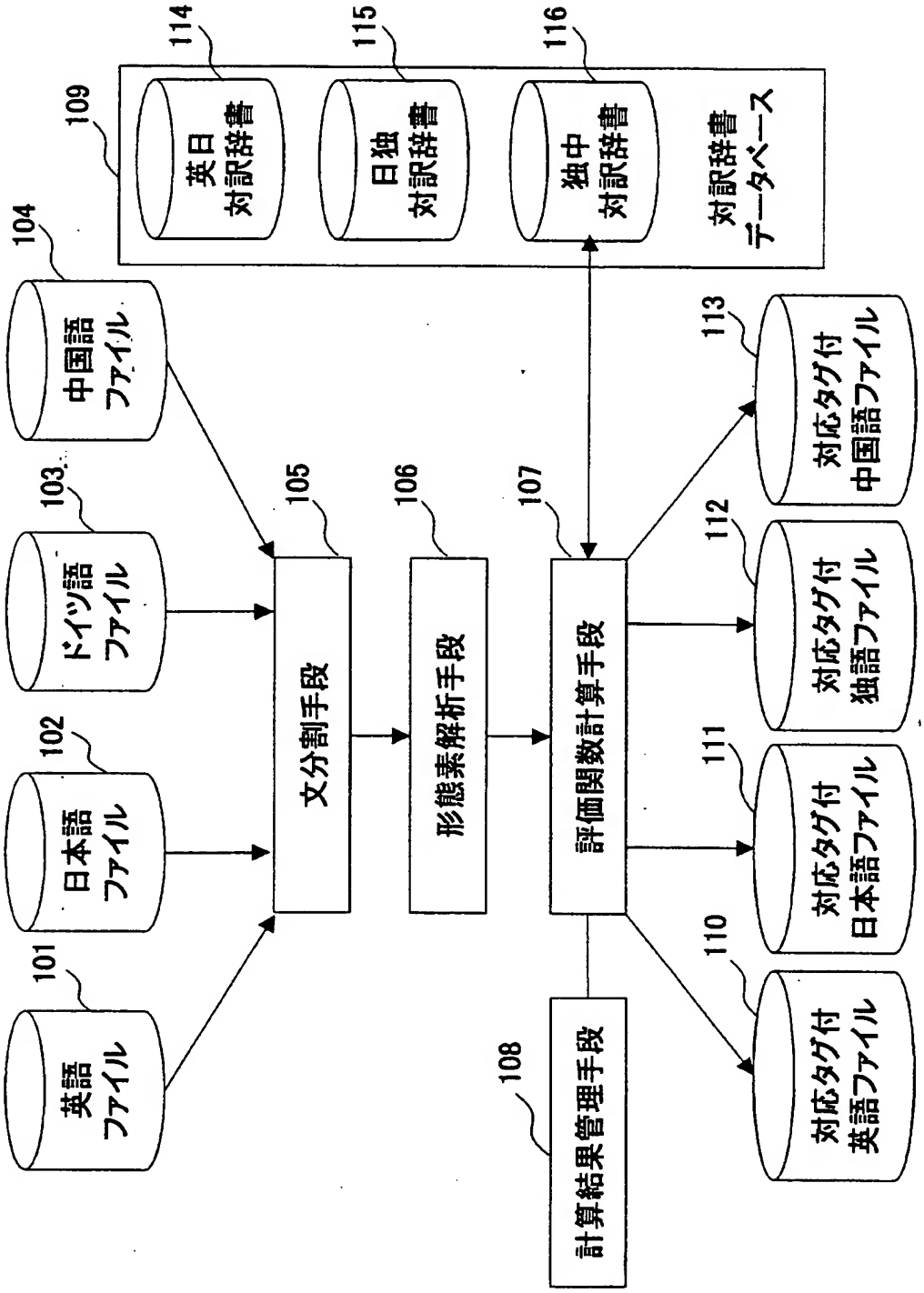
２１９，３１９，４１９　日中対訳辞書

２２０　相違箇所表示手段

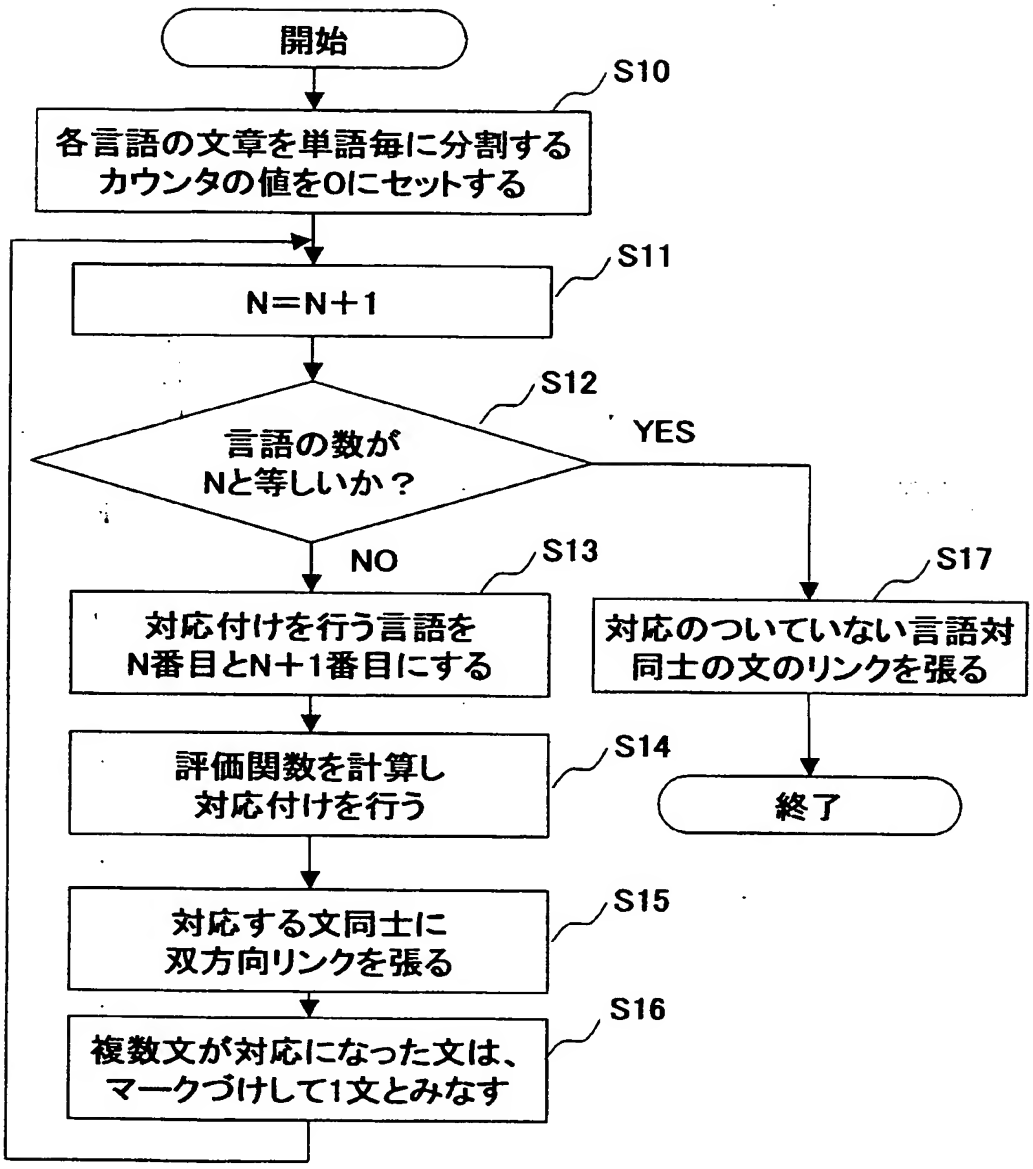
４２０　言語類似度データ

【書類名】　図面

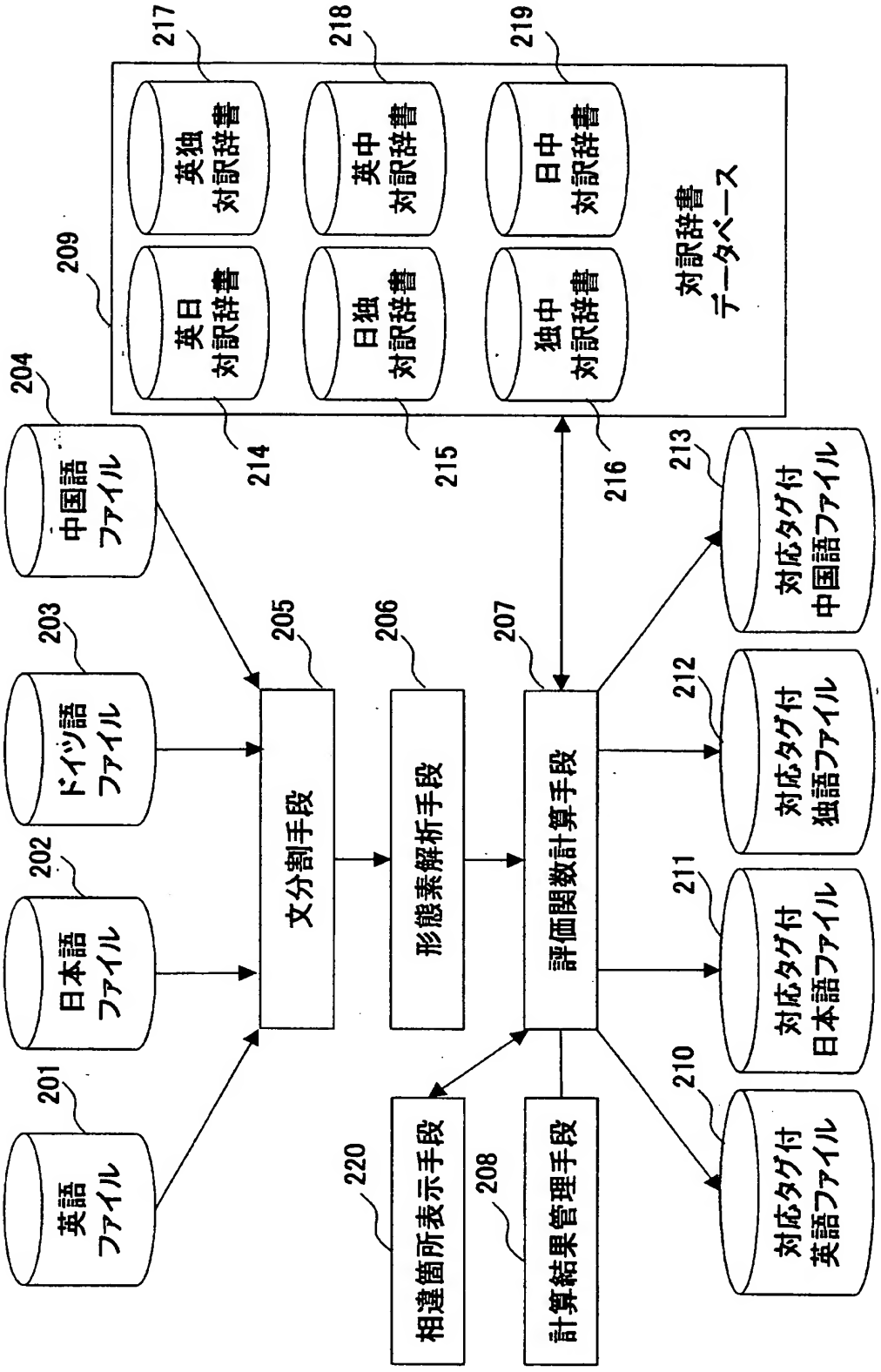
【図１】


114
115
116
109
英日
対訳辞書
日独
対訳辞書
独中
対訳辞書
対訳辞書
データベース
104
中国語
ファイル
103
ドイツ語
ファイル
102
日本語
ファイル
101
英語
ファイル
105
文分割手段
106
形態素解析手段
107
評価関数計算手段
108
計算結果管理手段
113
対応タグ付
中国語ファイル
112
対応タグ付
独語ファイル
111
対応タグ付
日本語ファイル
110
対応タグ付
英語ファイル

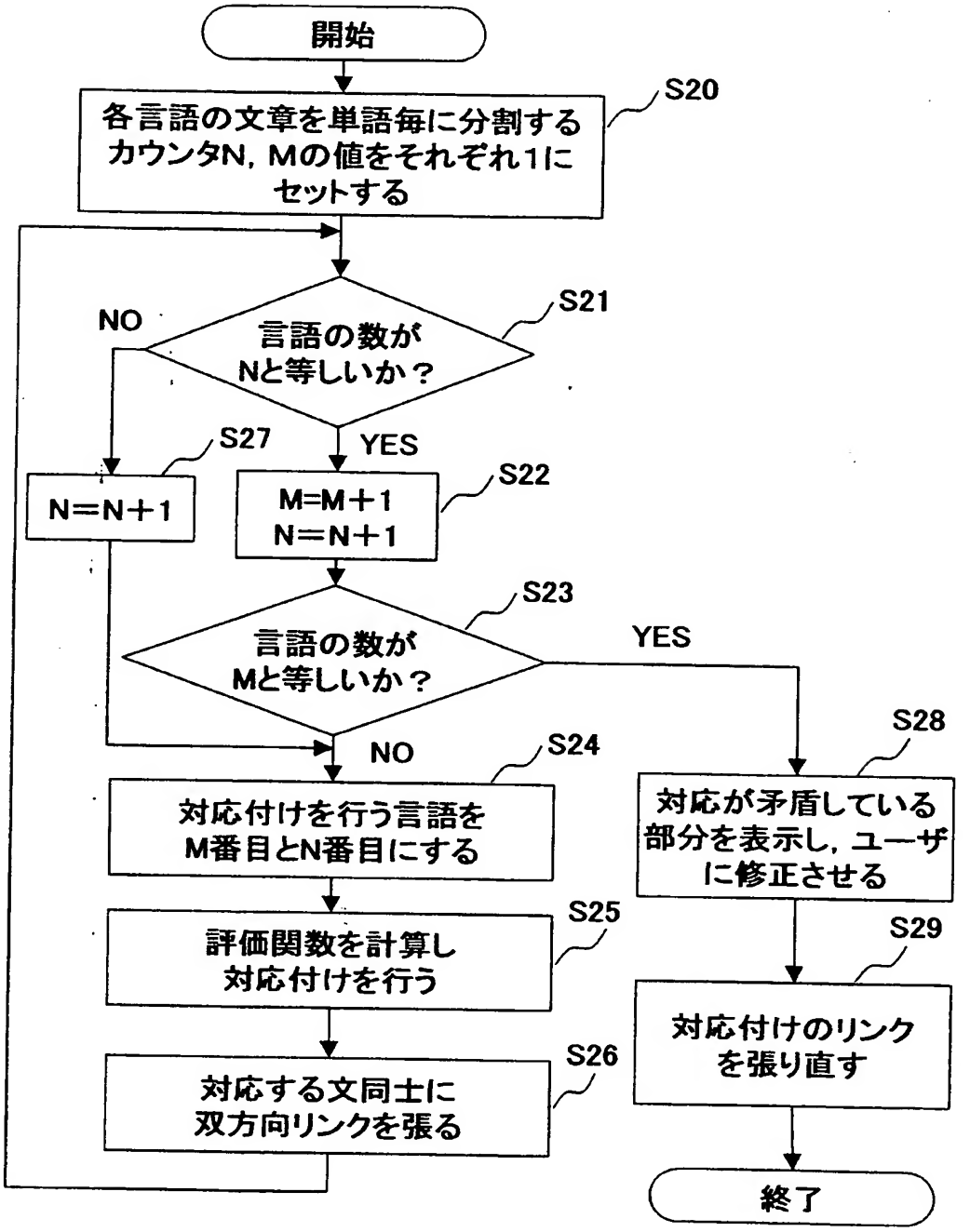
【図２】


開始
S10
各言語の文章を単語毎に分割する
カウンタの値をOにセットする
S11
N＝N＋1
S12
言語の数が
Nと等しいか？
YES
NO
S13
対応付けを行う言語を
N番目とN＋1番目にする
S14
評価関数を計算し
対応付けを行う
S15
対応する文同士に
双方向リンクを張る
S16
複数文が対応になった文は、
マークづけして1文とみなす
S17
対応のついていない言語対
同士の文のリンクを張る
終了

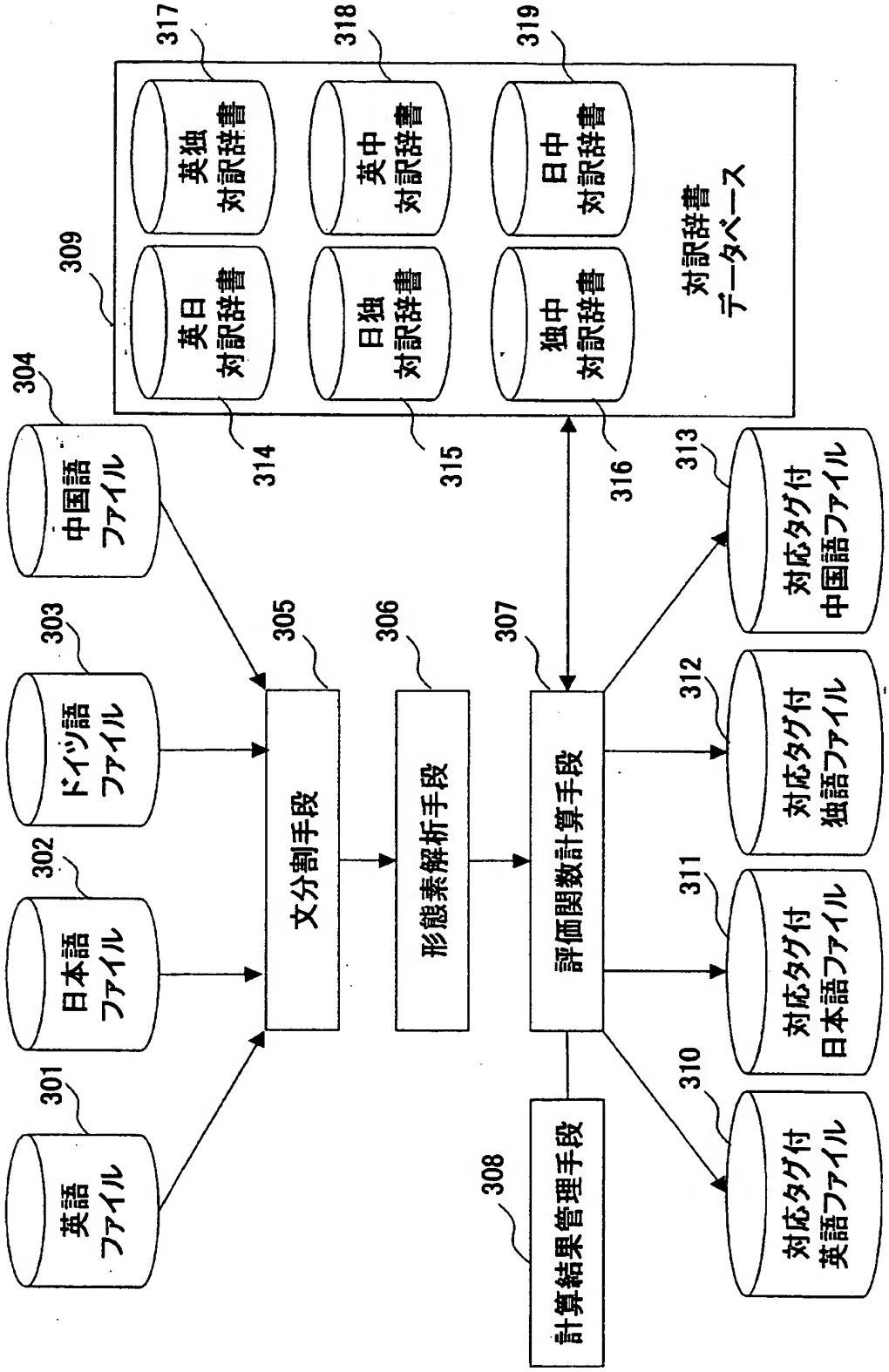
【図３】


201
英語ファイル
202
日本語ファイル
203
ドイツ語ファイル
204
中国語ファイル
205
文分割手段
206
形態素解析手段
207
評価関数計算手段
208
計算結果管理手段
209
対訳辞書データベース
210
対応タグ付英語ファイル
211
対応タグ付日本語ファイル
212
対応タグ付独語ファイル
213
対応タグ付中国語ファイル
214
英日対訳辞書
215
日独対訳辞書
216
独中対訳辞書
217
英独対訳辞書
218
英中対訳辞書
219
日中対訳辞書
220
相違箇所表示手段

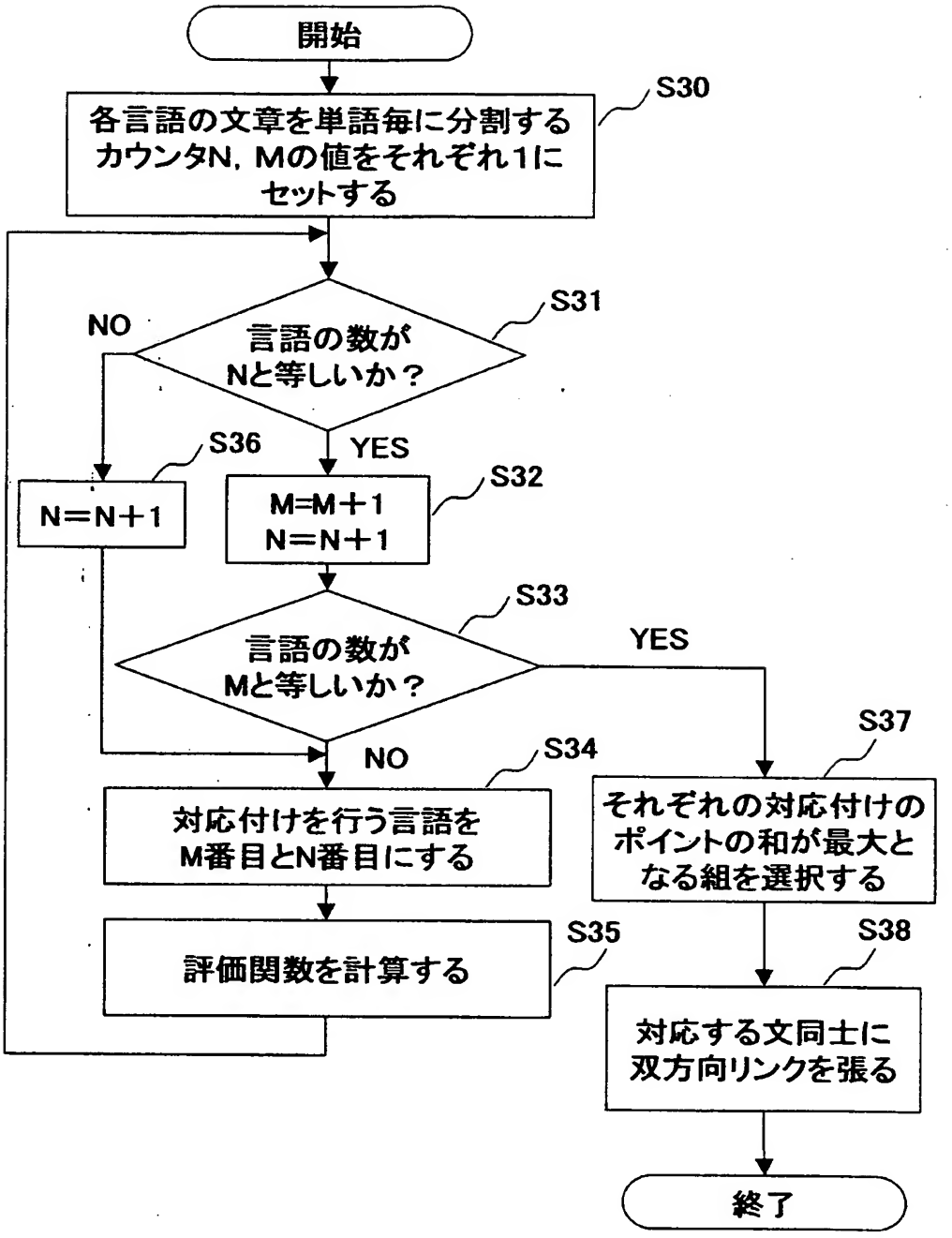
【図４】


開始
各言語の文章を単語毎に分割する
カウンタN，Mの値をそれぞれ1に
セットする
S20
言語の数が
Nと等しいか？
S21
NO
YES
N=N+1
S27
M=M+1
N=N+1
S22
言語の数が
Mと等しいか？
S23
YES
NO
対応付けを行う言語を
M番目とN番目にする
S24
対応が矛盾している
部分を表示し，ユーザ
に修正させる
S28
評価関数を計算し
対応付けを行う
S25
対応付けのリンク
を張り直す
S29
対応する文同士に
双方向リンクを張る
S26
終了

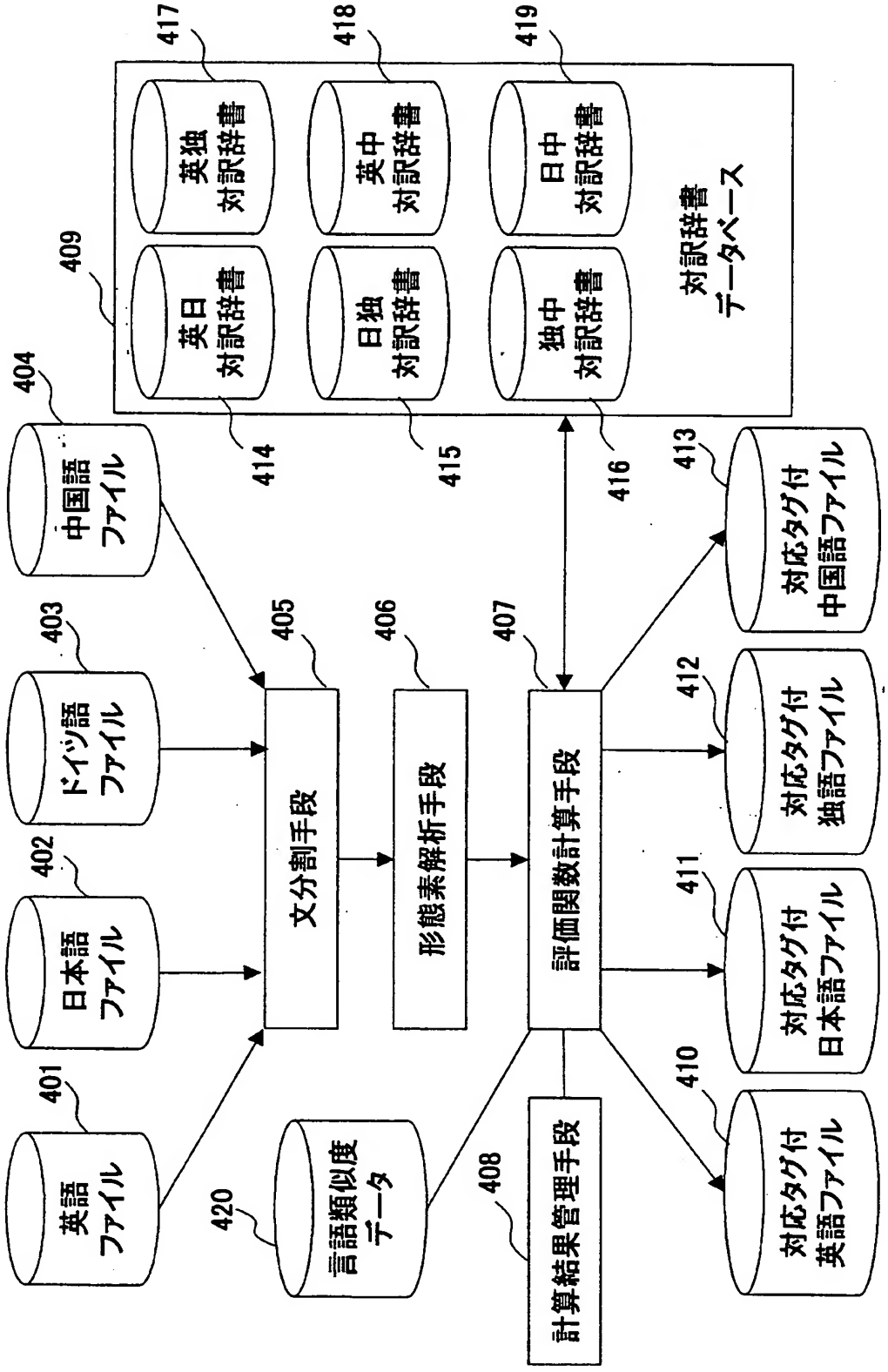
【図５】


317
318
319
英独
対訳辞書
英中
対訳辞書
日中
対訳辞書
309
英日
対訳辞書
日独
対訳辞書
独中
対訳辞書
対訳辞書
データベース
304
中国語
ファイル
314
315
316
313
対応タグ付
中国語ファイル
303
ドイツ語
ファイル
305
306
307
312
対応タグ付
独語ファイル
302
日本語
ファイル
文分割手段
形態素解析手段
評価関数計算手段
311
対応タグ付
日本語ファイル
301
英語
ファイル
310
対応タグ付
英語ファイル
308
計算結果管理手段

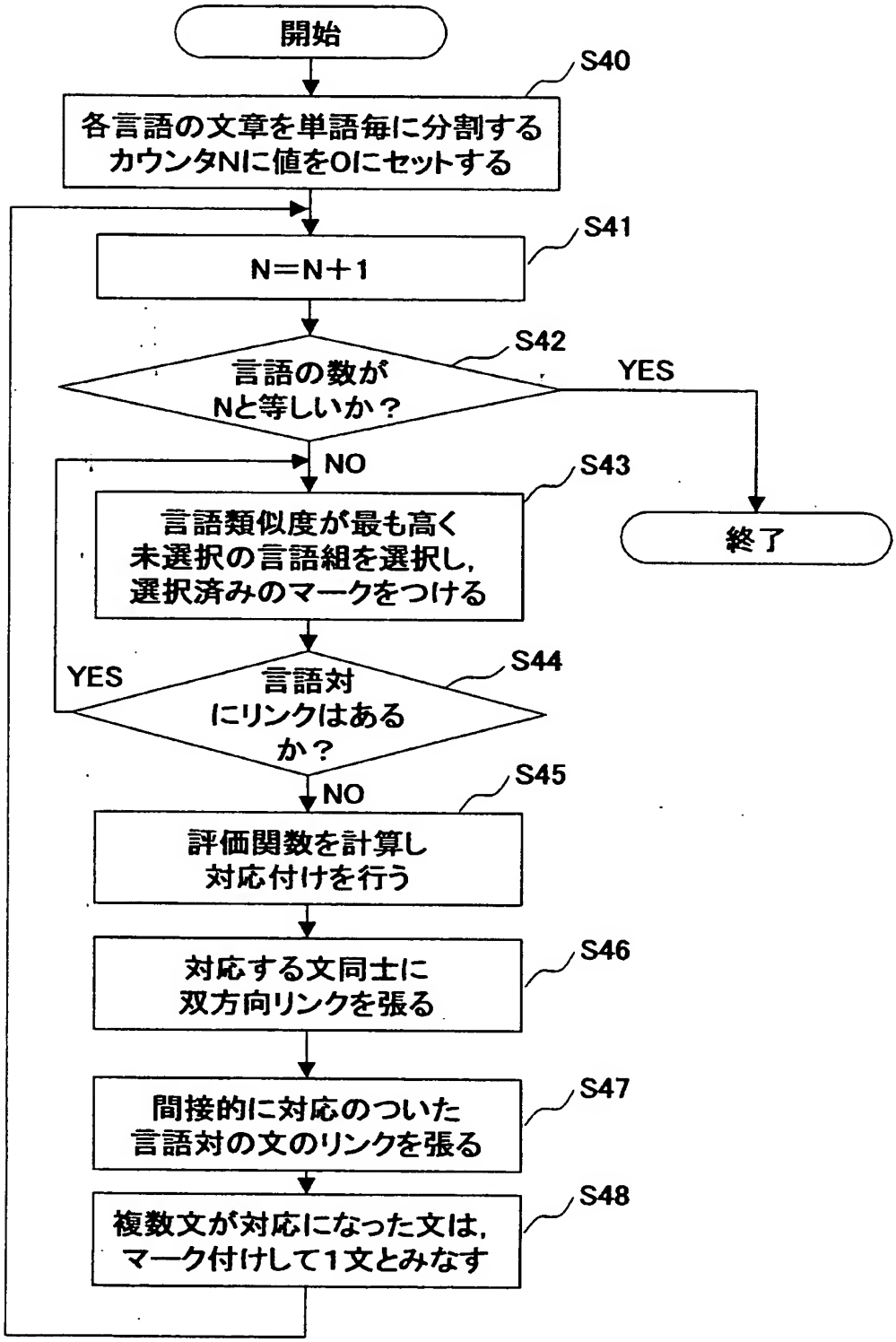
【図６】


開始
各言語の文章を単語毎に分割する
カウンタN，Mの値をそれぞれ1に
セットする
S30
言語の数が
Nと等しいか？
S31
NO
YES
S36
N＝N＋1
M=M＋1
N＝N＋1
S32
言語の数が
Mと等しいか？
S33
YES
NO
S34
対応付けを行う言語を
M番目とN番目にする
評価関数を計算する
S35
S37
それぞれの対応付けの
ポイントの和が最大と
なる組を選択する
S38
対応する文同士に
双方向リンクを張る
終了

【図７】


401
英語ファイル
402
日本語ファイル
403
ドイツ語ファイル
404
中国語ファイル
405
文分割手段
406
形態素解析手段
407
評価関数計算手段
408
計算結果管理手段
409
対訳辞書データベース
410
対応タグ付英語ファイル
411
対応タグ付日本語ファイル
412
対応タグ付独語ファイル
413
対応タグ付中国語ファイル
414
英日対訳辞書
415
日独対訳辞書
416
独中対訳辞書
417
英独対訳辞書
418
英中対訳辞書
419
日中対訳辞書
420
言語類似度データ

【図８】


開始
S40
各言語の文章を単語毎に分割する
カウンタNに値をOにセットする
S41
N＝N＋1
S42
言語の数が
Nと等しいか？
YES
NO
S43
言語類似度が最も高く
未選択の言語組を選択し，
選択済みのマークをつける
終了
YES
S44
言語対
にリンクはある
か？
NO
S45
評価関数を計算し
対応付けを行う
S46
対応する文同士に
双方向リンクを張る
S47
間接的に対応のついた
言語対の文のリンクを張る
S48
複数文が対応になった文は，
マーク付けして１文とみなす

【書類名】　要約書

【要約】

【課題】　複数の言語で構成される文書間の文の対応付けを効率良く行う複数言語文書の対応付けシステムを提供する。

【解決手段】　ｎ種（ｎは２以上）の言語の文書を単語毎に分割する形態素解析手段と，ｎ種の言語の文書のうちの２種を選択する手段と，選択された２種の言語文書の評価関数を計算する手段と，評価結果に応じてｎ種の言語の文書を対応付ける手段と，を含むことを特徴とする複数言語文書の対応付けシステム。各言語の文書を単語毎に分割する形態素解析手段は，各言語の文書を文毎に分割する手段と，分割された各文をさらに単語毎に分割する手段とからなってもよい。

【選択図】　図１

特願２００２－３４５９８８

出　願　人　履　歴　情　報

識別番号　　［０００００２９５］

１．変更年月日　　１９９０年　８月２２日
［変更理由］　　新規登録
住　所　　東京都港区虎ノ門１丁目７番１２号
氏　名　　沖電気工業株式会社